S. Kullback (1966): An Information-Theoretic Derivation of Certain Limit Relations for A Stationary Markov Chain. *SIAM Journal of Control*, August, pp. 454–459.

T. Lindvall (1992): *Lectures on the Coupling Method.* John Wiley and Sons Inc. , New York.

J.R. Norris (1997): *Markov Chains.* Cambridge University Press, Cambridge.

A.R. Raftery (1985): A Model for High-order Markov Chains. *Journal of the Royal Statistical Society*, B, 47, pp. 528 - 539.

Chapter 8

Learning of Markov Chains

8.1 Background

8.1.1 General Summary

Now we apply the steps of Bayesian modelling to a (training) sequence using a family of Markov models. A Markov model is completely specified by a transition matrix and an initial distribution. Probabilistic learning needs:

(1) a Markovian probability distribution which specifies the probability of any sequence conditioned by the transition matrix and the initial distribution;

(2) a prior which expresses the uncertainty about the transition matrix and the initial distribution.

When (1) is combined in a known fashion with the training sequence we obtain the likelihood function of the sequence with respect to a family of Markov models. The likelihood function is combined with (2) via Bayes' rule to produce a *posterior distribution* for the parameters of the family of Markov models. Using (1) and (2) we may also compute predictive distributions and do model comparison by means of Bayes factors. Model family comparison is especially concerned with finding the *order of the Markov chain*, as defined in Chapter 7, a technique appearing in modelling DNA sequences in Chapter 9.

8.2 ML for Markov Chains

8.2.1 Preliminaries

Let $\underline{\theta}$ be the transition probability matrix

$$\underline{\theta} = \begin{pmatrix} \theta_{1|1} & \theta_{1|2} & \dots & \theta_{1|J} \\ \theta_{2|1} & \theta_{2|2} & \dots & \theta_{2|J} \\ \vdots & \vdots & \vdots & \vdots \\ \theta_{J|1} & \theta_{J|2} & \dots & \theta_{J|J} \end{pmatrix}. \tag{2.1}$$

We are concerned with estimating the model $\underline{\theta}$ in the family of probabilistic models $p(\mathbf{x}|\underline{\theta})$ for a (training) sequence $\mathbf{x}$ of $n+1$ symbols in S^{n+1}

$$\mathbf{x} = (j_0 j_1 \dots j_n) \in S^{n+1}.$$

It was shown in Chapter 7 that for Markov chains

$$p(\mathbf{x}|\underline{\theta}) = P(X_0 = j_0, X_1 = j_1, \dots, X_n = j_n|\theta) = \pi_{j_0}(0) \prod_{l=1}^{n} \theta_{j_{l-1}|j_l}. \tag{2.2}$$

Then we propose

MODEL FAMILY:

CONDITIONED ON $\pi_{j_0}(0)$ AND $\Theta = \underline{\theta}$, THE SYMBOLS IN $\mathbf{x}$ ARE AN OUTCOME OF A MARKOV CHAIN $\{X_n\}_{n\geq 0}$, WITH STATIONARY TRANSITION PROBABILITIES $\underline{\theta}$. ▪

Here we have at most $J^2 - J$ transition parameters and the $J-1$ initial probabilities to estimate using the data $\mathbf{x}$. We make an approximation to the effect that we omit the initial distribution $\pi(0)$ as a part of the estimation problem. One way to think of this is that we know (or fix at will) the initial symbol in advance. Moreover, given just one training sequence we have just one single observation of the initial state. Let us first consider the maximum likelihood estimate of all of the unknown transition parameters.

As a function of $\underline{\theta}$ for fixed $\mathbf{x}$ the ensuing approximate or conditional likelihood function is

$$L(\underline{\theta}) = \prod_{l=1}^{n} \theta_{j_{l-1}|j_l}. \tag{2.3}$$

The corresponding log likelihood function is

$$\mathcal{L}(\underline{\theta}) = \sum_{l=1}^{n} \ln \theta_{j_{l-1}|j_l}. \tag{2.4}$$

8.2.2 ML of the Transition Matrix

We introduce a notation for the number of times we see a transition from i to j in $\mathbf{x} = (j_0 j_1 \dots j_n)$. Thus

$$n_{i|j} = \text{ the number of } l \text{ such that } 1 \leq l \leq n,\ j_{l-1} = i, j_l = j. \tag{2.5}$$

Using the frequency counts $n_{i|j}$ we can write the likelihood function as

$$L(\underline{\theta}) = \prod_{i=1}^{J} \prod_{j=1}^{J} \theta_{i|j}^{n_{i|j}}. \tag{2.6}$$

Obviously $n_{i|j}$s are the sufficient statistics for this model family. The log likelihood in (2.4) will be

$$\mathcal{L}(\underline{\theta}) = \sum_{i=1}^{J} \sum_{j=1}^{J} n_{i|j} \ln \theta_{i|j}. \tag{2.7}$$

Let also

$$n_i = \text{ the number of } l \text{ such that } 0 \leq l \leq n-1,\ j_l = i, \tag{2.8}$$

so that n_i is equal to the number of times the sequence $\mathbf{x}$ visits the state i, excluding the possible visit at the final time. Then we have

Proposition 8.2.1 *The maximum likelihood estimate $\widehat{\theta}_{i|j}$ of $\theta_{i|j}$ is*

$$\widehat{\theta}_{i|j} = \frac{n_{i|j}}{n_i}, \tag{2.9}$$

for all i and j.

Proof: Since the constraints

$$\theta_{i|j} \geq 0, \qquad \sum_{j=1}^{J} \theta_{i|j} = 1 \tag{2.10}$$

hold separately for each row in the transition matrix we can maximize $\mathcal{L}(\underline{\theta})$ in (2.7), which is a separable sum of the corresponding terms, by an independent maximization for each row. Thus for each row

$$\underline{\theta}_i = (\theta_{i|1}, \ldots, \theta_{i|J})$$

we should maximize

$$\mathcal{L}(\underline{\theta}_i) = \sum_{j=1}^{J} n_{i|j} \ln \theta_{i|j} \tag{2.11}$$

as a function of $\theta_{i|1}, \ldots, \theta_{i|J}$ so that the constraints are satisfied. Therefore we may repeat the computation from Chapter 3. Let us set

$$\widehat{\theta}_i = \left(\frac{n_{i|1}}{n_i}, \ldots, \frac{n_{i|J}}{n_i} \right).$$

Since $n_{i|1} + n_{i|2} + \ldots + n_{i|J} = n_i$, as every transition from i (possibly back to i) indicates necessarily a visit to i and the final time was excluded, we see that $\widehat{\theta}_i$ satisfies the constraints. Take now an arbitrary $\underline{\theta}_i$ satisfying the constraints. Then we get as in Chapter 3

$$\mathcal{L}\left(\widehat{\theta}_i\right) - \mathcal{L}(\underline{\theta}_i) = n_i D\left(\widehat{\theta}_i | \underline{\theta}_i\right) \geq 0,$$

where $D\left(\widehat{\theta}_i | \underline{\theta}_i\right)$ is the Kullback distance, which has been proved to be non-negative. Equality holds if and only if $\widehat{\theta}_i = \theta_i$. ∎

8.2.3 An Example of Full Likelihood

Suppose that the model family consists of *stationary* Markov chains with a binary state space S and with the transition probability matrices

$$A = \begin{pmatrix} 1-p & p \\ q & 1-q. \end{pmatrix}. \tag{2.12}$$

We wish now to estimate p and q using an observed sequence $\mathbf{x}$ of $n+1$ symbols in S^{n+1}

$$\mathbf{x} = (j_0 j_1 \ldots j_n) \in S^{n+1}.$$

In the preceding the initial distribution p_{X_0} was not a part of the estimation problem. If the chain is stationary then the initial distribution is an invariant distribution, which contains the unknown parameters. We can, of course, still throw away the initial distribution, but this means a loss of information, which is asymptotically insignificant.

The full likelihood function $L(p,q) = p(\mathbf{x}|A)$ for the stationary model given the sequence $\mathbf{x}$ turns out to be equal to

$$L(p,q) = \frac{p^a \cdot (1-p)^b \cdot q^c \cdot (1-q)^d}{p+q}, \tag{2.13}$$

where, using the notations for the number of state transitions in the sequence $\mathbf{x}$ introduced above,

$$a = j_0 + n_{0|1}, \qquad b = n_{0|0}, c = 1 - j_0 + n_{1|1}, \qquad d = n_{1|1}.$$

Hence there is no longer an explicit solution of the log likelihood equation obtained by setting the partial derivatives of $\ln L(p,q)$ equal to zero. In (Bisgaard and Travis 1991) it is shown that this system of equations has a unique solution which is a maximum.

8.3 The Whittle Distribution

For any given $\mathbf{x} = (j_0 j_1 \ldots j_n)$ we make the frequency counts (2.5) or

$$n_{i|j} = \text{ the number of } l \text{ such that } 1 \leq l \leq n,\ j_{l-1} = i, \qquad j_l = j.$$

Let us set

$$n_{i|\cdot} = \sum_{j=1}^{J} n_{i|j}, n_{\cdot|j} = \sum_{i=1}^{J} n_{i|j} \text{ for all } i \text{ and } j. \tag{3.1}$$

Thus $n_{i|\cdot}$ is the frequency count of i in the prefix $(j_0 j_1 \ldots j_{n-1})$ of $\mathbf{x}$ and $n_{\cdot|j}$ is the frequency count of j in the suffix $(j_1 \ldots j_n)$ and $\mathbf{x}$. Therefore

$$n_{i|\cdot} - n_{\cdot|i} = \delta_{ij_0} - \delta_{ij_n}, \tag{3.2}$$

where δ_{ij} is Kronecker's delta (i.e., $\delta_{ij} = 1$ if $i = j$ and $\delta_{ij} = 0$ if $i \neq j$) and

$$\sum^{J} n_{i|\cdot} = \sum^{J} n_{\cdot|j} = n. \tag{3.3}$$

Let

$$F = \left(n_{i|j}\right)_{i=1,j=1}^{J,J} \tag{3.4}$$

be any $J \times J$ matrix of non-negative integers which satisfy (3.2) and (3.3). Hence the knowledge of F and of j_0 determine j_n uniquely and F and j_n determine j_0.

Proposition 8.3.1 (Whittle's multinomial coefficient) *Let F be an $J \times J$ - matrix of non-negative integers $n_{i|j}$ such that $\sum_{i=1}^{J}\sum_{j=1}^{J} n_{i|j} = n$ and such that $n_{i|\cdot} - n_{\cdot|i} = \delta_{iu} - \delta_{iv}$ for some u and v in S. Let*

$N_{u,v}^{(n)}(F)$ = the number of sequences $\mathbf{x} = (j_0 j_1 \ldots j_n)$ having the frequency count F and satisfying $j_0 = u$, $j_n = v$.

Then

$$N_{u,v}^{(n)}(F) = \frac{\prod_{i=1}^{J} n_{i|\cdot}!}{\prod_{i=1}^{J}\prod_{j=1}^{J} n_{i|j}!} \cdot F_{vu}^{*}, \tag{3.5}$$

where F_{vu}^{*} is the (u,v)th cofactor of the matrix F^{*} with components

$$f_{ij}^{*} = \begin{cases} \delta_{ij} - \dfrac{n_{i|j}}{n_{i|\cdot}} & \text{if } n_{i|\cdot} > 0, \\ \delta_{ij} & \text{if } n_{i|\cdot} = 0. \end{cases}$$

Proof: A proof is given in (Billingsley 1962, pp. 14–15). The cofactor F_{vu}^{*} is $(-1)^{u+v}$ multiplied by the determinant of the matrix obtained obtained by deleting row u and column u, as defined in any text on matrices. ∎
Then we readily obtain t the probability that $\mathbf{x} = (j_0 j_1 \ldots j_n)$ has F as its transition count and $j_0 = u$ and $j_n = v$, denoted by $P_W(F)$, is (Whittle 1955) nothing else but

$$P_W(F) = \pi_u(0) \cdot F_{vu}^{*} \cdot \frac{\prod_{i=1}^{J} n_{i|\cdot}!}{\prod_{i=1}^{J}\prod_{j=1}^{J} n_{i|j}!} \cdot \prod_{i=1}^{J}\prod_{j=1}^{J} \theta_{i|j}^{n_{i|j}}. \tag{3.6}$$

A homogeneous Markov chain is thus seen to resemble a set of independent multinomial processes.

The Whittle distribution turns out to be useful in computing statistical properties of occurrences of words, a problem of considerable biological interest as shown in Chapter 9.

8.4 Model Averaging

8.4.1 Posterior Distributions for Rows in the Transition Matrix

Let us assume that our uncertainty about *the rows of $\underline{\theta}$* in (2.1)

$$\theta_i = \left(\theta_{i|1}, \ldots, \theta_{i|J}\right)$$

is modeled by *independent random variables* that have their respective Dirichlet densities for $i = 1, \ldots, J$ see section 3.8 in Chapter 3. These we formulate as

$$Dir\left(\theta_i; \alpha_i \cdot q_{i|1}, \ldots, \alpha_i \cdot q_{i|J}\right) = \frac{\Gamma(\alpha_i)}{\prod_{j=1}^{J}\Gamma\left(\alpha_i q_{i|j}\right)} \cdot \prod_{j=1}^{J} \theta_{i|j}^{\alpha_i q_{i|j}-1}, \tag{4.1}$$

where

$$\alpha_i > 0, \qquad q_{i|j} > 0, \qquad \sum_{j=1}^{J} q_{i|j} = 1.$$

Then we use as the simultaneous prior density the **multivariate Dirichlet density** or

$$\prod_{i=1}^{J} \mathrm{Dir}\left(\theta_i; \alpha_i \cdot q_{i|1}, \ldots, \alpha_i \cdot q_{i|J}\right) \tag{4.2}$$

suggested by (Martin 1967, ch. 2), see also (Basawa and Rao 1980 pp. 65 - 68). Hence the posterior density is in view of (2.3) equal to

$$p\left(\underline{\theta}|\mathbf{x}\right) = \frac{\prod_{i=1}^{J} \frac{\Gamma(\alpha_i)}{\prod_{j=1}^{J}\Gamma(\alpha_i q_{i|j})} \prod_{j=1}^{J} \theta_{i|j}^{n_{i|j}+\alpha_i q_{i|j}-1}}{p(\mathbf{x})}, \tag{4.3}$$

where $p(\mathbf{x})$ is the standardization that makes $p\left(\underline{\theta}|\mathbf{x}\right)$ a probability density in $\underline{\theta}$.

8.4.2 Predictive Probability

As an obvious extension of the predictive probabilities in Chapter 3 we might ask what is our probability

$$P\left(X_{n+1} = j | X_n = i; \mathbf{x}\right),$$

where the notation indicates that the probability is based on a given training sequence $\mathbf{x} = (j_0 j_1 \dots j_n) \in S^{n+1}$? In view of our Markov modelling of $\mathbf{x}$ in the preceding subsection one answer could be

$$\widehat{P}_{\mathrm{ML}}\left(X_{n+1} = j | X_n = i; \mathbf{x}\right) = \widehat{\theta}_{i|j}$$

plugging in the maximum likelihood estimate of the transition probability.

In a completely observabilistic sense we would consider only the single sequence $\mathbf{x}$ and ask for $P\left(X_{n+1} = j | \mathbf{x}\right)$ and provide the answer as

$$\widehat{P}_{\mathrm{ML}}\left(X_{n+1} = j | \mathbf{x}\right) = \widehat{\theta}_{j_n|j},$$

since j_n is the last symbol in the sequence.
There are other ways of addressing the stated question. Using the sequence $\mathbf{x}$ we may take some posterior density for $\underline{\theta}$ and then provide a new transition matrix by *model averaging*

$$P^*\left(X_{n+1} = j | X_n = i; \mathbf{x}\right) = \int \theta_{i|j} p\left(\underline{\theta} | \mathbf{x}\right) d\underline{\theta}. \tag{4.4}$$

Using (4.2)

$$\begin{aligned} P^*\left(X_{n+1} = r | X_n = s; \mathbf{x}\right) &= \int \theta_{r|s} p\left(\underline{\theta} | \mathbf{x}\right) d\underline{\theta} \\ &= \frac{\prod_{i=1}^{J} \frac{\Gamma(\alpha_i)}{\prod_{j=1}^{J} \Gamma\left(\alpha_j q_{i|j}\right)} I_{i|j}(r,s)}{p(\mathbf{x})}, \end{aligned}$$

where

$$I_{i|j}(r,s) = \int \prod_{j=1}^{J} \theta_{r|s} \theta_{i|j}^{n_{i|j}\alpha_i q_{i|j} - 1} d\theta_i.$$

Here the integration is with respect to all of the parameters in $\underline{\theta}$, which are, however, separated to their respective domains as rows of $\underline{\theta}$. Using the well known formulae for evaluating the various Dirichlet integrals (appendix to Chapter 3 and Chapter 6) we obtain in (4.4) the expression

$$P^*\left(X_{n+1} = r | X_n = s; \mathbf{x}\right) = \frac{n_{s|r} + \alpha_s q_{s|r}}{n_s + \alpha_s}. \tag{4.5}$$

The parameters α_s and $q_{s|r}$ play, as before, the role of pseudo counts of observations or of regularizers.

8.5 MC Order Comparison Using the Bayes Ratio

In the literature on biological sequence analysis the problem of estimating the order of a time-homogeneous Markov chain and/or testing against a null model, the multinomial process with the same state space. We now state a relevant criterion, without restricting ourselves to any specific biological situation, using Bayesian model comparison. We compute the Bayes ratio

$$B(\mathbf{x}) = \frac{q_M(\mathbf{x})}{q_{M_0}(\mathbf{x})}, \tag{5.1}$$

where under the model family M the training sequence $\mathbf{x}$ is related to the parameters in a transition matrix $\underline{\theta}$ as above and with the multivariate Dirichlet density (4.2) as prior. Under the model family M_0 the training sequence $\mathbf{x}$ is related to the parameters in a conditional independence model with a Dirichlet prior (cf. Chapter 3).

As in the chapter quoted we obtain

$$q_{M_0}(\mathbf{x}) = \frac{\Gamma(\alpha)}{\Gamma\left(\prod_{i=1}^{J} \alpha q_i\right)} \cdot \frac{\prod_{i=1}^{J} \Gamma\left(\alpha q_i + n_i^*\right)}{\Gamma(n + \alpha)}, \tag{5.2}$$

where n_i^* is equal to the number of times the symbol i appears in $\mathbf{x}$. Note that by the definitions valid here n_i^* is equal to n_i for some $J-1$ symbols and $n_i + 1$ for the remaining of them. In view of the formulas above we have that

$$\begin{aligned} q_M(\mathbf{x}) &= \prod_{i=1}^{J} \frac{\Gamma(\alpha_i)}{\prod_{j=1}^{J} \Gamma\left(\alpha_i q_{i|j}\right)} \int_{\underline{\theta}} \prod_{j=1}^{J} \theta_{i|j}^{n_{i|j} + \alpha_i q_{i|j} - 1} d\underline{\theta} \\ &= \prod_{i=1}^{J} \frac{\Gamma(\alpha_i)}{\prod_{j=1}^{J} \Gamma\left(\alpha_i q_{i|j}\right)} \cdot \frac{\prod_{j=1}^{J} \Gamma\left(n_{i|j} + \alpha_i q_{i|j}\right)}{\Gamma\left(n_i + \alpha_i\right)}. \end{aligned} \tag{5.3}$$

The idea in (Milosaljevic and Jurka 1993) can be recapitulated as searching a database for sequences $\mathbf{x}$ such that $-\log B(\mathbf{x})$ exceeds a threshold. The threshold can be taken as the length of the codeword for $\mathbf{x}$ compressed by a suitable algorithm, an application of the theory in Chapter 2